Searcher in a Strange Land: Understanding Web Search from Familiar and Unfamiliar Locations

Elad Kravi[1,2], Eugene Agichtein[2,4], Ido Guy[2], Yaron Kanza[3], Avihai Mejer[2], Dan Pelleg[2]
[1] Technion, Israel [2] Yahoo Labs, Israel [3] Cornell Tech, USA [4] Emory University, USA
ekravi@cs.technion.ac.il, {agichtein,amejer}@yahoo-inc.com, {idoguy,pellegd}@acm.org, kanza@cornell.edu

ABSTRACT

With mobile devices, web search is no longer limited to specific locations. People conduct search from practically anywhere, including at home, at work, when traveling and when on vacation. How should this influence search tools and web services? In this paper, we argue that information needs are affected by the familiarity of the environment. To formalize this idea, we propose a new contextualization model for activities on the web. The model distinguishes between a search from a familiar place ($\mathcal{F}$*-search*) and a search from an unfamiliar place ($\mathcal{U}$*-search*). We formalize the notion of familiarity, and propose a method to identify familiar places. An analysis of a query log of millions of users, demonstrates the differences between search activities in familiar and in unfamiliar locations. Our novel take on search contextualization has the potential to improve web applications, such as query autocompletion and search personalization.

Keywords

Web search modeling; search personalization; query language modeling; location-based search

1. INTRODUCTION

Mobile devices have led to a paradigm shift from web search that is mainly conducted in specific locations to search that is being done virtually anywhere and is frequently related to the geographic environment. Consequently, many search engines provide a location-centric search, where the answer is affected by the location of the user, and it has been shown that this improves search results [1]. However, searches by residents and searches by infrequent visitors typically reflect different needs. For example, a search query "train dog", issued by a visitor, should refer to the regulations of traveling with a dog on a train. But for a local, it should relate to puppy training. As another example, in New York City, an obvious query auto-completion of "natural" is "natural history museum" for a tourist, but "natural gas" or "natural food" for residents.

To address this problem, we propose a new approach to geographical personalization—an approach that is *user*-centric and is based on how familiar the searcher is with the location. Previous studies along this line explored a functional typology, and in particular focused on a small set of labels ("home", "work", and sometimes "school") [10]. We argue that a complementary view revolves around the level of familiarity users have with their surroundings. In this respect, "home" would be considered a familiar place (as would "work" or "school"), *e.g.*, in contrast to a city that the user visits for the first time. The difference being that in a familiar place, basic needs of the user, such as food, transportation and place to live in, are generally part of a routine. In these familiar environments, constituting a user's "natural habitat", the user is likely to have prior knowledge for many basic needs, especially those related to the physical environment. As a result, in a familiar environment, the user may be free to focus on higher-level search activities that satisfy leisure needs and/or may require deeper level of involvement. We call a search in such environments $\mathcal{F}$*-search*, and argue that these environments result in a distinct set of information needs and search patterns.

The natural complement, still under the user-centric view, are unfamiliar places. This generally corresponds to an out-of-town scenario, but not always—a frequented location is considered "familiar", even if distant from the user's home. An unfamiliar place is simply a place that is seldom visited, or is visited for the first time, and we refer to a search in such places as $\mathcal{U}$*-search*.

In this work, we formally define a model that distinguishes between $\mathcal{F}$-search and $\mathcal{U}$-search and we apply it to a search query log on a large scale. We hypothesize that the information needs in unfamiliar locations are different from those in familiar locations and we show experimentally that familiarity affects the type of web searches users conduct.

Several studies compared information needs in mobile versus desktop search [2–4, 6]. Our distinction between search from familiar and unfamiliar locations is device-independent: while we expect more searches from mobile devices to be from unfamiliar locations, our model does not explicitly consider the device type and our findings suggest that both types of searches are conducted from both desktop and mobile devices. Yom-Tov and Diaz [9] showed that in a case of a natural disaster, the information needs considerably change based on the distance from the event's location. White and Buscher [8] compared searches of restaurants by local and non-local users and showed that local knowledge enabled finding venues of higher quality. None of these studies, how-

SIGIR '15, August 09–13, 2015, Santiago, Chile.
Copyright 2015 ACM 978-1-4503-3621-5/15/08 ...$15.00.
http://dx.doi.org/10.1145/2766462.2767782.

ever, propose or study the distinction between familiar and unfamiliar search locations.

The contributions of this paper include the following.

- A new characterization of search queries to distinguish between $\mathcal{F}$-search in "familiar" places versus $\mathcal{U}$-search in "unfamiliar" locations, defined on a per-user basis.
- Analysis of massive search logs, showing that the distinction provides useful information.
- Qualitative analysis that demonstrates the differences in language models and suggests potential applications, such as query auto-completion.

2. CONTEXTUALIZATION MODEL

We now formally define search familiarity. Let U be a set of users. Each user $u \in U$ is associated with a set of search activities S_u, where an activity is a query submitted to a search engine followed by a click on one of the results.

To each search activity s in S_u, we associate the time of the search and the location of the user at that time, denoted $t(s)$ and $l(s)$, respectively. An activity is also associated with the search query, the clicked link, and the type of the device on which the search was conducted.

The *search locations* of a user u is the set $L_u = \{l(s) \mid s \in S_u\}$ of all the locations of the activities of u. Our goal is to distinguish in L_u between familiar locations, in which u is regularly active, and the locations in which the activity is sporadic. In this research, we partition the geographical space into areas, according to zip-codes, and consider each area as a *place*. Each place is associated with the activities whose location is in this place, and for a specific user u and a place P, the set $S_u(P)$ is the set of searches of u within P.

When the user conducts many searches merely from a single place, we naturally assume that this place is familiar to the user. Next, we consider cases where the user performs searches from more than one place. Intuitively, we define a place as familiar if the search activities of the user reveal that the user has been active in this place during many days and in several occasions.

Formally, we define $Days_u(P)$ as the set of days on which u conducted a search from P, *i.e.*, all the days that contain a time $t(s)$ for some search activity $s \in S_u$ in place P. By $Days_u$, we denote all the days on which there is some search activity of u (in any place). We say that the *time spent* of u at P is the ratio $TS_u(P) = \frac{|Days_u(P)|}{|Days_u|}$. A *return* of user u to place P is a case where first, there is a search activity of u in P, then a search activity of u from at least one place different than P, and again, a search from P.

Definition 2.1 ($\mathcal{F}$-search). *A place P is familiar to user u if one of the following two cases holds. (i) All the activities of u are in P. (ii) Given a time-spent ratio t and a return count r, P is familiar with respect to t and r, if* (1) *$TS_u(P) \geq t$, i.e. the time spent of u at P is at least t, and* (2) *the number of returns of u to P is at least r. A search activity of user u in a familiar place of u is considered an $\mathcal{F}$-search.*

A place P that is not familiar to user u is considered *unfamiliar* to u and we refer to a search activity of u at such place as *$\mathcal{U}$-search*.

In our experiments, we opted to set r to 2, requiring that a user will return at least twice to a place in order for it to become a familiar place for the user (unless it is the sole search place). Parameter t was set to be 10%, because this balanced between the two conditions, giving them a roughly equal weight. That is, inspecting all pairs (u, P), such that user u conducted a search activity in place P, applying merely Condition 1 with $t = 10\%$ deemed 53.6% of the places in these pairs as familiar, while applying merely Condition 2 with $r = 2$ deemed 53.3% of the places as familiar, causing both conditions to be almost equally selective. With both of these parameters jointly applied, 51.4% of the places were considered familiar to their corresponding user.

3. SEARCH CHARACTERISTICS

In this section, we analyze the different aspects of search activities in familiar and unfamiliar places. First, we present our experimental setup and provide some statistics regarding users and their searches.

Our dataset includes more than a billion search actions sampled from the logs of a popular commercial search engine during a period of six months. These queries were posted by more than 35 million unique users from about 30,000 US zip codes. The dataset includes only users with at least 20 recorded search activities and at least 20 days of activity. On average, during the inspected time period, each of these users performed search activities from 9.15 different places (stdev: 12.79, median: 4). For places, the average number of users per place was 3887 (stdev: 5123.91, median 235).

Distribution of time spent: Obviously, search activities are not conducted everywhere and all the time. So, how are they distributed? There are many places in which people are only seldom active and only a few places in which people are active during many days. To illustrate this, we calculated the distribution of activity days per places and the results are depicted in Fig. 1. A point (d, p) on the graph indicates that in p percent of the search places, the user was active during at least d percent of the days. While 69.3% of the places account for at least 1% of the activity days, only 19.2% of the places account for at least 5% of the days; 11.0% account for at least 10% of the days, and in 6.4% of the places, the activity spanned more than 20% of the user's activity days. This shows the effect of changing t in Definition 2.1—increasing t will reduce the number of places that are deemed familiar, roughly following a power-law distribution.

Home as a familiar place: As a first step of verifying our model, we examined how often the defined home of a user u is deemed a familiar place of u. To this end, we define the *minimum $\mathcal{F}$-search distance from home* as the minimal distance between the location of an $\mathcal{F}$-search and the location of the declared home of the user (for users with at least one familiar place and a declared home). Fig. 2 shows the percentage of users whose minimum $\mathcal{F}$-search distance from home was smaller than a given distance d. For 53.9% of these users, the distance was smaller than 20 kilometers and for 75.4% of the users it was smaller than 100 kilometers. This result agrees with prior work which analyzed all page views, rather than just search actions [7]. Overall, this confirms that in many cases the home of the user is a familiar place, but not always, since the declared home is not always the actual home of the user (e.g., a fake or obsolete address).

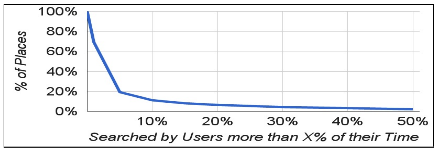

Figure 1: Distribution of activity time in places.

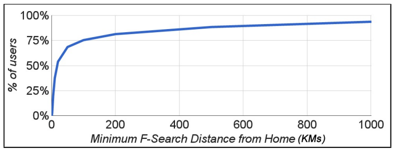

Figure 2: Distribution of the minimum $\mathcal{F}$-search distance from home.

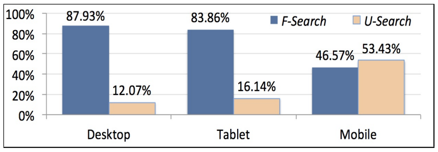

Figure 3: Percentage of $\mathcal{F}$-search and $\mathcal{U}$-search for different devices.

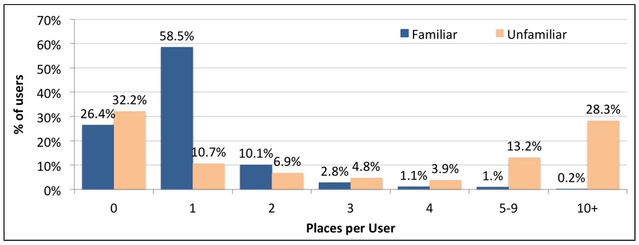

Figure 4: Distribution of the number of familiar and unfamiliar places per user.

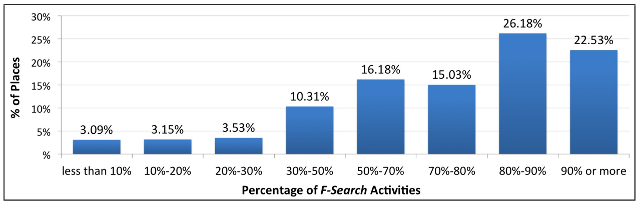

Figure 5: $\mathcal{F}$-search portion per place.

Familiar and unfamiliar places: Fig. 4 presents the distribution of the number of familiar and unfamiliar places per user, according to Definition 2.1. About a quarter of the users do not have familiar places at all; the majority of the users (58.5%) have one familiar place, while a little over 15% of the users have two or more familiar places. As for unfamiliar places, while 32.2% of the users have no $\mathcal{U}$-search activities at all, for the remaining users, the number of unfamiliar places in which they were active is typically much higher than the number of familiar places: only 10.7% of the users have exactly one unfamiliar place in which they posed a query, while 28.3% of the users have ten or more unfamiliar places and 13.7% have 20 or more unfamiliar places.

$\mathcal{F}$-Search and $\mathcal{U}$-search activities per place: Fig. 5 presents the distribution of the portion of $\mathcal{F}$-searches per place. In only 3.1% of the places, less than 10% of the searches were $\mathcal{F}$-searches. In almost 80% of the places (16.2+15.0+26.2+22.5), there are more $\mathcal{F}$-searches than $\mathcal{U}$-searches. In nearly 50% of the places (26.2 + 22.5), over 80% of the search activities were $\mathcal{F}$-searches.

The effect of the device: In Fig. 3, we see that the device affects the percentage of $\mathcal{F}$-searches. As expected, on desktops and on tablets, almost all the searches are $\mathcal{F}$-searches, whereas on mobile phones, slightly more than 50% of the searches are $\mathcal{U}$-searches. Surprisingly, in this measure, tablets are more similar to desktops than to mobile phones.

$\mathcal{F}$-Search vs. $\mathcal{U}$-search query language: To gain better intuition into the differences between $\mathcal{F}$-searches and $\mathcal{U}$-searches, we investigated which keywords characterize queries sent from familiar places as compared to queries sent from unfamiliar places. To this end, we built two specialized language models (LMs)—one for $\mathcal{F}$-searches and one for $\mathcal{U}$-searches. To build language models that would scale to the large query log volume, we used the Berkeley LM library [5], which provides an efficient and scalable implementation of n-gram LMs. For unknown terms we used standard Laplace smoothing with ϵ of 1 over the size of the vocabulary in the training data. After creating the LMs, we set out to identify the terms with the highest divergence between the familiar and unfamiliar models. Specifically, Table 1 presents the top 12 uni-grams and bi-grams with the highest Kullback-Leibler (KL) divergence scores between the two models, after stop-word removal.

Uni-grams		Bi-grams	
$\mathcal{F}$-Search	$\mathcal{U}$-Search	$\mathcal{F}$-Search	$\mathcal{U}$-Search
facebook	google	for sale	new york
sale	restaurant	how to	phone number
free	schedule	facebook login	google search
games	football	to make	new jersey
ebay	ny	homes for	high school
how	lyrics	cool math	how many
login	ct	you tube	hobby lobby
online	store	sales in	in new
craiglist	movie	funeral home	football schedule
recipes	hours	real estate	r us
porn	locations	black friday	movie theater
tube	mall	for kids	nfl scores

Table 1: The top 12 distinctive query terms of $\mathcal{F}$-searches and $\mathcal{U}$-searches.

It can be seen that the $\mathcal{F}$-search lists consist of activities related to shopping, social networks, games, knowledge seeking and adult content. The $\mathcal{U}$-search lists, in contrast, are focused on searching for location or schedule of restaurants, movies, malls, cities, institutes, and sport events. The top word in the uni-gram lists reflect the fundamental difference between the two types of information needs: it is "google" for $\mathcal{U}$-searches and "facebook" for $\mathcal{F}$-searches. Both are the two most popular sites on the web (according to Alexa.com). The first is focused on search and the second on social networking, indicating general information needs while in unfamiliar places, compared to stronger social information needs in familiar places.

Inspecting the main categories of the top 100 uni-grams and bi-grams, we found similar trends. The top $\mathcal{F}$-search queries included keywords that relate to the following categories: `shopping` (15%), e.g., 'sales' and 'amazon'; `social media` (10%), e.g., 'twitter', 'tumblr' and 'youtube'; `multi-media` (9%), e.g., 'video', 'photos', 'netflix', 'watch'; `adult content` (6%), e.g., 'porn', 'pornhub', 'sex'; `news` (6%), e.g., 'news', 'yahoo', and 'msn'; `email` (5%), e.g., 'gmail', 'hotmail'; `games` (4%), e.g., 'angry birds', 'games'; `family` (4%), e.g., 'wife', 'children'; `healthcare` (4%), e.g., 'medicine', 'symptoms', 'pregnant'.

Category	Prefix	Familiar	Unfamiliar
Food	coffee	table:2, shop:4	shop:2, table:5
	steak	marinade:1, house:2	house:1, marinade:4
	pizza	dough:3, places:5	places:3, dough:5
Travel	gas	fireplace:3, station:6	station:3, fireplace:8
	train	crash:1, schedule:7	schedule:2, crash:6
	car	games:2, wash:6	wash:2, games:7
Local	court	cases:1, house:4	house:1, cases:2
	science	news:4, museum:10	museum:3, news:7
	police	scanner:2, department:9	department:2, scanner:4
Leisure	wild	rice:2, horse:5	horse:2, rice:4
	soccer	drills:4, scores:13	scores:5, drills:7
	piano	sheet:2, bar:8	bar:2, sheet:4

Table 2: Sample word completion patterns reporting the first word (prefix), and example completions, excluding stop words, with rank, sorted by decreasing completion frequency for the familiar and unfamiliar settings, respectively.

The top $\mathcal{U}$-searches included keywords related to the following categories: `entertainment` (25%), e.g., 'restaurant', 'movie', 'mall', 'club'; `tourism` (16%), e.g., 'airport', 'weather', 'los angeles', 'texas'; `navigation` (15%), e.g., 'directions', 'location', 'street'; `sports` (5%), e.g., 'nfl', 'espn', 'scores'.

Overall, in familiar places, people search for more profound information that requires time (and state-of-mind) exploring, watching, reading and interacting. In unfamiliar places, on the other hand, there is a greater need for quick lookups for more "touristic" information needs such as food and entertainment, with focus on navigation, directions, and opening times.

4. DISCUSSION

We defined and demonstrated some of the differences between $\mathcal{F}$-search and $\mathcal{U}$-search. This distinction can be of value in practical applications for personalization of search results, news-feed items, or recommendations. Specifically, we now demonstrate how $\mathcal{F}$-search/$\mathcal{U}$-search distinction can be useful for query auto-completion.

4.1 Query Auto-Completion

In query auto-completion, the user types a few characters or the first word of the search query, and the system predicts potential queries. Based on the observations above and on our analysis of familiar and unfamiliar query patterns, we propose to use this information to improve query auto-completion. Specifically, we propose to use different completions for familiar versus unfamiliar places. To gain some intuition into the different completions, we present examples of different completions for common query categories in Table 2. For these examples, we report the initial word of the query, in the dataset described above, together with excerpts from the query completion lists, ranked in decreasing order by completion frequency (in the format of '<completion word>:<rank>'). For example, in the "Food" category, the $\mathcal{F}$-searches focus on aspects more likely to be performed in one's home, such as marinading a steak or buying a coffee table. In contrast, in unfamiliar places, food aspects more likely to focus on immediate needs, such as pizza places (higher on the list for unfamiliar places) or steak houses. A similar effect can be observed in the "Travel", "Local", and "Leisure" categories. For example, for the term "train", the completion "crash" is the first ranked in familiar places, while less frequent (6^{th}) at unfamiliar places, whereas "schedule" is ranked high (2^{nd}) in unfamiliar places, but lower (7^{th}) in familiar places; for the term "court", "cases" is the top completion in familiar places, as users may have more time for looking into these, while "house" is the top completion in unfamiliar places, reflecting more ad-hoc needs. These results suggest that for popular query categories, search suggestions ranking should take user's location familiarity into account.

4.2 Other Applications

Different applications can benefit from a new contextualization model that distinguishes between familiar and unfamiliar locations of search. Obviously, and based on our preliminary results, in web search it is beneficial to provide different results to $\mathcal{F}$-searches and to $\mathcal{U}$-searches. A natural extension is to use the proposed distinction in web advertising and recommender systems: for unfamiliar places, touristic services would be suggested, but in $\mathcal{F}$-search, daily needs will populate the ads and recommendations. Even navigation systems may benefit from the suggested distinction—in an unfamiliar surrounding, people prefer the simplest routes to follow. But locals will take the fastest (traffic-adjusted) way, however convoluted. We are exploring these topics for follow-up works.

5. REFERENCES

[1] P. N. Bennett, F. Radlinski, R. W. White, and E. Yilmaz. Inferring and using location metadata to personalize web search. In *Proc. of SIGIR*, 2011.

[2] K. Church and B. Smyth. Understanding the Intent Behind Mobile Information Needs. In *International Conf. on Intelligent User Interfaces*, 2009.

[3] M. Kamvar and S. Baluja. A Large Scale Study of Wireless Search Behavior: Google Mobile Search. In *Proc. of SIGCHI*, 2006.

[4] M. Kamvar, M. Kellar, R. Patel, and Y. Xu. Computers and Iphones and Mobile Phones, Oh My!: A Logs-based Comparison of Search Users on Different Devices. In *Proc. of WWW*, 2009.

[5] A. Pauls and D. Klein. Faster and smaller n-gram language models. In *Proc. of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, 2011.

[6] D. Pelleg, D. Savenkov, and E. Agichtein. Touch Screens for Touchy Issues: Analysis of Accessing Sensitive Information from Mobile Devices. In *ICWSM*, 2013.

[7] D. Pelleg, E. Yom-Tov, and Y. Maarek. Can you believe an anonymous contributor? on truthfulness in Yahoo! answers. In *SocialCom*, 2012.

[8] R. White and G. Buscher. Characterizing Local Interests and Local Knowledge. In *SIGCHI*, 2012.

[9] E. Yom-Tov and F. Diaz. Out of Sight, Not out of Mind: On the Effect of Social and Physical Detachment on Information Need. In *SIGIR*, 2011.

[10] H. Zang and J. Bolot. Anonymization of location data does not work: A large-scale measurement study. In *Proc. of MobiCom*, 2011.